
Journal of Urban Health: Bulletin of the New York Academy of Medicine Vol. 79, No. 2, June 2002

Power for *T*-Test Comparisons of Unbalanced Cluster Exposure Studies

Donald R. Hoover

ABSTRACT *Studies of individuals sampled in unbalanced clusters have become common in health services and epidemiological research, but available tools for power/sample size estimation and optimal design are currently limited. This paper presents and illustrates power estimation formulas for* t*-test comparisons of effect of an exposure at the cluster level on continuous outcomes in unbalanced studies with unequal numbers of clusters and/or unequal numbers of subjects per cluster in each exposure arm. Iterative application of these power formulas obtains minimal sample size needed and/or minimal detectable difference. SAS subroutines to implement these algorithms are given in the Appendices. When feasible, power is optimized by having the same number of clusters in each arm* $k_A = k_B$ *and (irrespective of numbers of clusters in each arm) the same total number of subjects in each arm* $n_A k_A = n_B k_B$*. Cost beneficial upper limits for numbers of subjects per cluster may be approximately* $(5/\rho) - 5$ *or less where* ρ *is the intraclass correlation. The methods presented here for simple cluster designs may be extended to some settings involving complex hierarchical weighted cluster samples.*

KEYWORDS *Cluster Sampling, Power, Sample Size,* T *Tests, Unbalanced Designs.*

INTRODUCTION

Cluster studies sample or treat individuals in clusters.[1] Often, all individuals in the cluster have the same exposure (or treatment) status, and the mean outcome of the cluster has a normal or approximately normal distribution. Since exposure is the same for all individuals in a cluster, this is denoted as a "cluster exposure" in this article. We focus on simple (one-level) unweighted cluster studies of a binary exposure with all individuals in the cluster having the same exposure status. Extensions to hierarchical weighted cluster samples are described in Appendix 3.

Consider two examples of simple cluster exposure studies. For example 1, in a study of effect of neighborhood crime on mental health status in New York City, 10 individuals may be sampled from each of 400 neighborhoods (the cluster), of which 100 are "high crime exposure" and 300 are "low crime exposure" neighborhoods. Mean mental health status would be compared between the high crime exposure and low crime exposure neighborhoods. For example 2, in a study of the effect of support group size (10 vs. 30 individuals) on behavioral intervention to reduce sexually transmitted disease (STD) spread, 300 individuals may be randomly placed into support groups of size 10 (for a total of 30 groups), while another 300 are randomly placed into 10 support groups of size 30. Postintervention mean sex-

Dr. Hoover is with the Department of Statistics, Rutgers University, 473 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019. (E-mail: drhoover@stat.rutgers.edu)

ual risk behavior score will be compared between each of these two support group size arms (cf. Susser et al.[2]).

When sampling (or intervening) in clusters, subjects from the same cluster can be positively correlated with each other. Hence, the cluster is the unit of analysis, and the cluster effect must be adjusted.[3,4] While this depends on the nature of the clustering, intracluster correlation ρ tends to be higher for smaller clusters,[3] perhaps as individuals in smaller clusters have more contact with each other. In most applications,[4,5] ρ is no greater than 0.6. Standard formulas for estimation of study power and sample size that require observations to be independent[4,6] cannot be used for cluster studies with $\rho > 0$. Power and sample size estimates have been presented for cluster studies with equal numbers of clusters and the same number of subjects per cluster in both arms.[4,7–9] However, frequently in epidemiologic and health outcomes research (such as examples 1 and 2 above), cluster studies will be unbalanced, with either the number of clusters and/or the number of individuals per cluster differing between comparison arms. We present power formulas for such unbalanced cluster exposure designs.

NOTATION AND ASSUMPTIONS

Assume k_A clusters in exposure arm A and n_A subjects in each cluster of arm A. Similarly, assume k_B clusters in exposure arm B and n_B subjects per cluster of arm B. The total number of subjects is $N = k_A(n_A) + k_B(n_B)$. When $n_A = n_B$, the common value is denoted n. Let X_{Aij} and X_{Bij} be the jth observation from the ith cluster in arms A and B, respectively. Because of the cluster effect, there are two components of variance for each observation: $X_{Aij} = u_A + \gamma_{Ai} + \varepsilon_{Aij}$, where u_A is the overall mean for arm A, γ_{Ai} is a random effect of cluster i from arm A with $\gamma_{Ai} \sim N(o,\sigma_i^2)$,and ε_{Aij} is the random effect of the jth individual in the ith cluster of arm A with $\varepsilon_{Aij} \sim N(o,\sigma_j^2)$. Similarly, $X_{Bij} = u_B + \gamma_{Bi} + \varepsilon_{Bij}$, with u_B the overall mean for arm B, $\gamma_{Bi} \sim N(o,\sigma_i^2)$ a random effect of cluster i from arm B and $\varepsilon_{Bij} \sim N(o,\sigma_j^2)$ the random effect of the jth individual in the ith cluster of arm B. All random components are independent.

The overall variance of an observation is $\sigma^2 = \sigma_i^2 + \sigma_j^2$, and the correlation of two observations from the same cluster $\rho = \sigma_i^2/(\sigma_i^2 + \sigma_j^2)$ is between 0 and 1. Thus, $\overline{X}_{A,i} = \sum_{j=1}^{n_a} X_{Aij}/n_A$ and $\overline{X}_{B,i} = \sum_{j=1}^{n_B} X_{Bij}/n_B$ (the means of the ith clusters in arms A and B, respectively) have expectations μ_A and μ_B, respectively, and variances $\sigma^2(1/n_A + \rho(1 - 1/n_A))$ and $\sigma^2(1/n_B + \rho(1 - 1/n_B))$, respectively, or $\sigma^2(1/n + \rho(1 - 1/n))$ for both when the number of subjects per cluster is equal for both arms.

We test the null hypothesis $H_o: \mu_A = \mu_B$ either against a two-sided alternative $H_a: \mu_A \neq \mu_B$ or against a one-sided alternative $H_a: \mu_A < \mu_B$ or $H_a: \mu_a > \mu_B$ using a two-sample t test. A more complicated approach to make this comparison would be a nested one-way analysis of variance (ANOVA) or hierarchical model that weighted cluster means by cluster size.[10,11] For example 1, $n_A = n_B$ (see the section "Power With Equal Group Sizes but Unequal Numbers of Clusters per Arm"), the nested ANOVA and two-sample t tests give identical results. For example 2, $n_A \neq n_B$ (see the section, "Power for Studies With Unequal Group Sizes and/or Numbers of Clusters per Arm"), the standard nested ANOVA model is not valid.[12] While a Satterthwaite[13] approach would create a more unbiased nested ANOVA test, this would be hard to implement in practice and, unlike the two-sample t test, is not robust to cluster exposure arms having different intraclass correlations.

TWO-SAMPLE *T*-TEST STATISTICS FOR COMPARISONS OF CLUSTERS

The test statistic t_r is constructed from the cluster means as the clusters are the smallest independent units: If variances of the cluster means are equal for clusters from exposure arms A and B, (i.e., when $n_A = n_B = n$), then

$$t_r = \frac{\overline{X}_A - \overline{X}_B}{S_p\sqrt{1/k_A + 1/k_B}}$$

where

$$\overline{X}_A = \frac{\sum \overline{X}_{A,i}}{k_A},\ \overline{X}_B = \frac{\sum \overline{X}_{B,i}}{k_B}$$

$$S_p = \sqrt{\frac{\sum(\overline{X}_{A,i} - X_A)^2 + \sum(\overline{X}_{B,i} - X_B)^2}{(k_A - 1) + (k_B - 1)}} \qquad \text{and} \qquad r = k_A + k_B - 2. \tag{1}$$

The test statistic t_r from Eq. 1 has a central t distribution with r degrees of freedom (df).[6] In Eq. 1, S_p^2/k_A estimates the variance of $\overline{X}_A$ and S_p^2/k_B the variance of $\overline{X}_B$, so $S_p\sqrt{1/k_A + 1/k_B}$ estimates the standard deviation of $\overline{X}_A - \overline{X}_B$. As variances of cluster means in arms A and B are equal, we use a combined estimator S_P^2.

Variances of the cluster means differ for clusters in exposure arms A and B when $n_A \neq n_B$, thus the test statistic t_r is

$$t_r = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{S_A^2/k_A + S_B^2/k_B}}$$

where

$$\overline{X}_A = \frac{\sum \overline{X}_{i,A}}{k_A},\ S_A = \sqrt{\frac{\sum(\overline{X}_{i,A} - \overline{X}_A)^2}{k_A - 1}},\ \overline{X}_B = \frac{\sum \overline{X}_{i,B}}{k_B}$$

$$S_B = \sqrt{\frac{\sum(\overline{X}_{i,B} - \overline{X}_B)^2}{k_B - 1}} \qquad \text{and} \qquad r = \frac{(S_A^2/k_A + S_B^2/k_B)^2}{(S_A^4/(k_A^2(k_A - 1)) + S_B^4/(k_B^2(k_B - 1)))}. \tag{2}$$

By Satterthwaite's[13] method, the test statistic t_r from Eq. 2 has an approximate central t distribution with r degrees of freedom. As cluster mean variance differs between arms A and B, S_A^2/k_A estimates the variance of $\overline{X}_A$, S_B^2/k_B estimates the variance of $\overline{X}_B$, and $\sqrt{S_A^2/k_A + S_B^2/k_B}$ estimates the standard deviation of $\overline{X}_A - \overline{X}_B$.

For both Eq. 1 and Eq. 2, the df, r reflect imprecision in t_r resulting from the fact that estimates of the cluster mean standard deviations (S_p, S_A, and S_B) rather than the true standard deviations being used. Estimates are more precise for larger values of r. For $r > 120$, both Eq. 1 and Eq. 2 have approximate standard normal distributions.[6]

Figure 1 gives a diagram of the of the distribution of the test statistic in either Eq. 1 or Eq. 2 and rejection regions chosen to make the (two-sided) type I error equal to $\alpha = .05$; $r = 22$ *df* is used for illustrative purposes.

We refer to the lower and upper t values associated with the rejection as $t_r(\alpha/2)$ and $t_r(1 - \alpha/2)$, respectively. Since central t distributions are symmetric about zero, $t_r(\alpha/2) = -t_r(1 - \alpha/2)$, and we reject the null hypothesis at a given type I error α when the absolute value of the test statistic from Eq. 1 or Eq. 2 exceeds $t_r(1 - \alpha/2)$. Making α smaller increases the value of $t_r(1 - \alpha/2)$ needed to provide an overall type I error α. In Fig. 1, the numerical value of $t_{22}(1 - 0.05/2)$ is 2.074. For one-sided testing with an alternative hypothesis, rejection of $H_o: \mu_A > \mu_B$ occurs if the test statistic exceeds $t_r(1 - \alpha)$.

BASIC FORMULA FOR POWER

The power of a study for a specified level of the alternative hypothesis is defined as $1 - \beta$, where β is the type II error or, equivalently, the probability the test statistic will fall into the appropriate rejection region if the alternative hypothesis is true at a specified level. A minimal significant value Δ (or a likely value for Δ) is chosen such that $\mu_A - \mu_B = \Delta$ under the specified level of H_a. Without loss of generality, let $\Delta > 0$ and the appropriate rejection region be greater than $t_{1-\alpha/2}$; we could reverse the labels of A and B if Δ was negative. If this level of H_a is true, then the test statistic has a noncentral t distribution with r degrees of freedom (the same value of r as for the distribution of the test statistic in Eqs. 1 and 2 under the null hypothesis) and a noncentrality parameter Φ where $\Phi = \Delta/\text{Std Dev}\ (\overline{X}_A - \overline{X}_B)$. The mean of

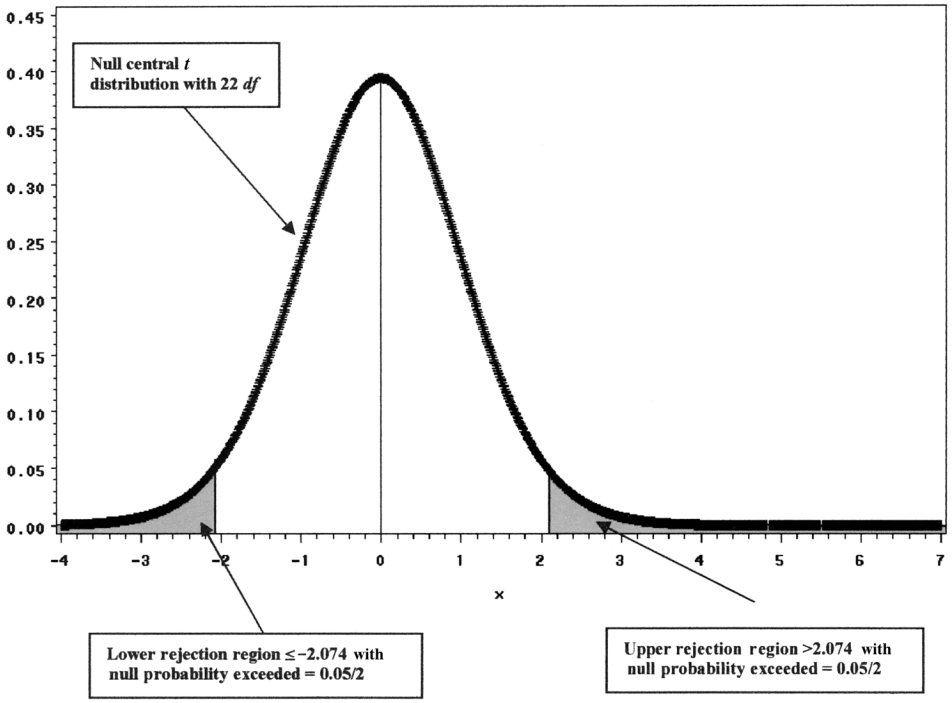

FIGURE 1. Rejection region for a t test with $r = 22$ *df* and two-sided $\alpha = .05$.

a noncentral t variable is approximately equal to Φ; this Φ indicates how many standard deviations of the test statistic the alternative hypothesis is from the null hypothesis. Figure 2 shows the rejection region of the null hypothesis for $df = 22$ and $\alpha = .05$ along with the probability for the test statistic to fall into the upper rejection region when $\Phi = 3$, or the alternative mean is three standard deviations of the test statistic greater than the null hypothesis mean of zero.

Remember that $2.074 = t_{22}(1 - .05/2)$, which defines the two-sided rejection region with an overall type I error of 0.05. If the specified alternative hypothesis is true with Φ [or Δ/Std Dev $(\overline{X}_A - \overline{X}_B)$] equal to 3, then the probability the test statistic does not exceed the rejection value of 2.074 (or the type II error under the specified alternative hypothesis) is the probability that a noncentral t distribution with 22 df and noncentrality parameter 3 is less than 2.074. We denote this probability as $t^{-1}_{22,3}$ (2.074), where 22 is df, 3 is the noncentrality parameter, and the inverse superscript (t^{-1}) in $t^{-1}_{22,3}(2.074)$ is the probability that a noncentral t distribution with $df = 22$ and $\Phi = 3$ is 2.074 or less. As Fig. 2 shows, for a given type I error α, $df(r)$, minimal significant value for $H_a: \Delta$, and Std Dev$(\overline{X}_A - \overline{X}_B)$, the power to detect the alternative hypothesis is

$$1 - \beta = 1 - t^{-1}_{r,\Phi}[t_r(1 - \alpha/2)] \qquad (\text{or } 1 - t^{-1}_{r,\Phi}[t_r(1 - \alpha)] \text{ for one sided testing}). \tag{3}$$

where $\Phi = \Delta$/Std Dev$(\overline{X}_A - \overline{X}_B)$. From examining Fig. 2, everything else being equal: (1) decreasing α increases $t_r(1 - \alpha/2)$, which makes the formula of Eq. 3 smaller; (2)

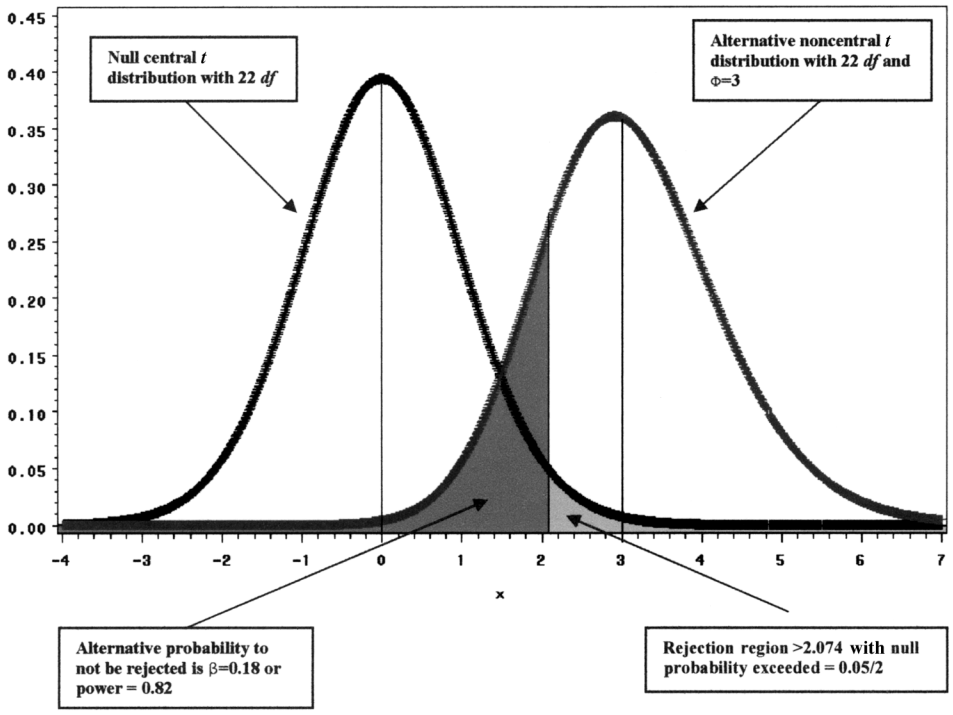

FIGURE 2. Rejection region for a t test with $r = 22$ df, $\alpha = .05$, and corresponding type II error under alternative hypothesis with noncentrality parameter $\Phi = 3$.

increasing Δ increases Φ, which makes Eq. 3 larger; and (3) increasing Std Dev($\overline{X}_A - \overline{X}_B$) decreases Φ, which makes Eq. 3 smaller. Increasing r reduces the size of the tails for central and noncentral t distributions, which often (but not always) makes Eq. 3 larger.

As the shape of a noncentral t distribution is similar to that of a central t distribution with the same degrees of freedom, Eq. 3 can be approximated by $1 - \beta = t_r^{-1}[t_r(1 - \alpha/2) - \Phi]$. For $r > 120$, t distributions can also be approximated by the normal distribution, and Eq. 3 is equivalent to $Z^{-1}[\Phi - Z_{1-\alpha/2}]$, where $Z^{-1}[\cdot]$ denotes the probability a standard normal variable is less than $[\cdot]$. In the examples used in this paper, powers from normal and central t approximations to the noncentral t distribution were within 0.02 of the exact powers from Eq. 3, although deviations will be greater for smaller r.

For calculation of power of the unbalanced cluster designs described in the introduction, Eq. 3 can be directly applied with appropriate values of r and Φ according to the study design, as described in the next two sections.

POWER WITH EQUAL GROUP SIZES BUT UNEQUAL NUMBERS OF CLUSTERS PER ARM

In the setting of power with equal group sizes but unequal numbers of clusters per arm, $\mathrm{Var}(\overline{X}_{A,i}) \equiv \mathrm{Var}(\overline{X}_{B,i}) = \sigma^2(1/n + \rho(1 - 1/n))$ and Std Dev$(\overline{X}_A - \overline{X}_B) = \sigma\sqrt{(1/k_A + 1/k_B)(1/n + \rho(1 - 1/n))}$. Thus, for a given alternative $\mu_A - \mu_B = \Delta$, the noncentrality parameter and degrees of freedom, respectively, are

$$\Phi = \Delta/\sigma\sqrt{(1/k_A + 1/k_B)(1/n + \rho(1 - 1/n))} \qquad \text{and} \qquad r = k_A + k_B - 2. \tag{4}$$

Placing these values of Φ and r into Eq. 3 gives, for two-sided testing,

$$1 - \beta = 1 - t^{-1}_{k_A+k_B-2,\,\Delta/\sigma\sqrt{(1/k_A+1/k_B)(1/n+\rho(1-1/n))}}[t_{k_A+k_B-2}(1 - \alpha/2)]. \tag{5}$$

Appendix 1 contains documented SAS (SAS Institute, Cary, NC) code to implement the power calculation from Eq. 5. The larger ρ is, the smaller Φ in Eq. 4 becomes, thus greater intracluster dependency reduces the noncentrality parameter and corresponding power. Everything else being equal, increasing n, k_A, and/or k_B increases the noncentrality parameter in Eq. 4 and power. For a fixed number of clusters $k_T = k_A + k_B$, the sum $(1/k_A + k_B)$ is minimized, and thus the noncentrality parameter in Eq. 4 is maximized when $k_A = k_B$. For a fixed N, larger values of k_A and k_B will increase r and increase Φ and thus increase power.

To illustrate computation of power using Eq. 5, return to example 1, with $k_A = 100$ high-crime and $k_B = 300$ low-crime neighborhoods in New York City; $n = 10$ persons sampled from neighborhoods in both arms. Suppose the outcome of interest, mental health function, has a standard deviation of $\sigma = 20$, intraneighborhood correlation of $\rho = 0.40$, and an overall mean of 50. We wish to test the null hypothesis with a two-sided $\alpha = .05$ and be able to detect a difference of $\Delta = 5$ between the mean scores of persons in high-crime neighborhoods and those in low-crime neighborhoods.

In this example, illustrated in Fig. 3, $\Phi = \Delta/\sigma\sqrt{(1/k_A + 1/k_B)(1/n + \rho(1 - 1/n))} = 5/20\sqrt{(1/100 + 1/300)(1/10 + .04(1 - 1/10))} = 3.19$, $r = k_A + k_B - 2 = 100 + 300 - 2 = 398$, and the rejection region $[t_{k_A+k_B-2}[1 - \alpha/2)]$ is $[t_{398}(1 - .025)] = 1.96$. From

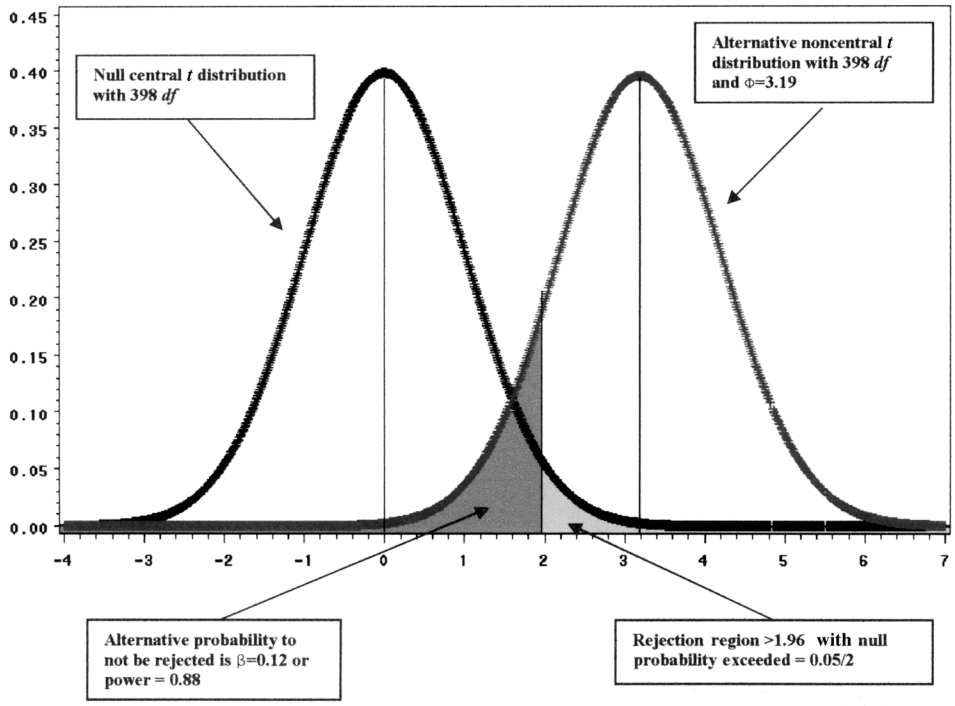

FIGURE 3. Rejection region for t test of example 1 with $r = 398$ *df* and $\alpha = .05$ and corresponding type II error under alternative hypothesis with noncentrality parameter $\Phi = 3.19 = 5/20\sqrt{(1/100 + 1/300)(1/10 + 0.4(1 - 1/10))}$

Eq. 5 then, $1 - \beta = 1 - t^{-1}_{398,3.19}[1.96]$ which since 398 is greater than 120, also equals $Z^{-1}[3.19 - 1.96] = Z^{-1}[1.20] = 0.88$. This study has 88% power to detect $\Delta = 5$. By contrast, a naïve power estimate based on the false assumption that all individuals in the same exposure cluster were independent (i.e., ignoring the intraclass correlation) would falsely estimate the power to be virtually 1.

POWER FOR STUDIES WITH UNEQUAL GROUP SIZES AND/OR NUMBERS OF CLUSTERS PER ARM

In the setting of power for studies with unequal group sizes and/or numbers of clusters per arm, since $n_A \neq n_B$, the variances of the cluster means are not equal; $\mathrm{Var}(\overline{X}_{Ai}) = \sigma^2(1/n_A + \rho(1 - 1/n_A))$ and $\mathrm{Var}(\overline{X}_{Bi}) = \sigma^2(1/n_B + \rho(1 - 1/n_B))$. Because of these unequal variances, Eq. 2 is used to test the null hypothesis. The test statistic theoretically does not have an exact t distribution under either the null or alternative hypothesis, but in both settings can be approximated by a t distribution.[13] Di Santostefano and Muller[14] found that a close approximation to the power obtains from Eq. 3 with noncentrality parameter $\tilde{\Phi} = \Delta/\sqrt{U_A + U_B}$ where $U_A = \sigma^2(1/n_A + \rho(1 - 1/n_A))/k_A$ (the variance of $\overline{X}_A$), $U_B = \sigma^2(1/n_B + \rho(1 - 1/n_B))/k_B$ (the variance of $\overline{X}_B$), and degrees of freedom

$$\tilde{r} = \left[(U_A^2)\frac{k_A + 1}{k_A - 1} + 2U_AU_B + (U_B^2)\frac{k_B + 1}{k_B - 1} \right] \Bigg/ \left[(U_a^2)\frac{k_A + 1}{(k_A - 1)^2} + (U_B^2)\frac{k_B + 1}{(k_B - 1)^2} \right]. \quad (6)$$

Briefly, $\tilde{\Phi}$ is Δ/Std Dev$(\overline{X}_A - \overline{X}_B)$, a direct implementation of Eq. 3, while $\tilde{r}$ is the expectation of $\dfrac{(S_A^2/k_A + S_B^2/k_B)^2}{(S_A^4/(k_A^2(k_A-1)) + S_B^4/(k_B^2(k_B-1)))}$ given that S_A and S_B are random variables. With $\tilde{\Phi}$ and $\tilde{r}$ incorporated, the power is approximately

$$1 - \beta = 1 - t^{-1}_{\tilde{r},\tilde{\Phi}}[t_{\tilde{r}}(1 - \alpha/2)]. \tag{7}$$

Appendix 2 contains a documented SAS program to implement the power calculation in Eq. 7. While patterns are complicated, in general power is maximized when $U_A = U_B$. If feasible, for a fixed N and $k_A = k_B$, one obtains better power with $n_A = n_B$ and using Eq. 2 versus $n_A \neq n_B$ with Eq. 3. In other words, if the number of exposure clusters is equal in both arms (and assuming the costs of individuals in each arm is the same) from a power standpoint, it is optimal to have the number of subjects in the clusters of each arm be equal.

To illustrate computation of power from Eq. 7, return to example 2, the postintervention mean risk behavior score comparison between 300 subjects randomized into $k_A = 30$ support cluster groups of exposure size $n_A = 10$ subjects and 300 different subjects randomly placed into $k_B = 10$ support cluster groups of exposure size $n_B = 30$ subjects each. From analysis of the data from Susser et al.,[2] the standard deviation of the sexual activity score is about one unit, a difference of $\Delta = 0.30$ units corresponds to a 50% reduction in human immunodeficiency virus (HIV) transmission activities, and the correlation of subjects in the same support group is $\rho = 0.30$. Then,

$$U_A = \sigma^2(1/n_A + \rho(1 - 1/n_A))/k_A = 1(1/10 + .3(1 - 1/10))/30) = 0.0123$$
$$U_B = \sigma^2(1/n_B + \rho(1 - 1/n_B))/k_B = (1(1/30 + .3(1 - 1/30))/10) = 0.0323$$
$$\tilde{\Phi} = \Delta/\sqrt{U_A + U_B} = 0.30/\sqrt{0.0123 + 0.0323} = 1.42$$

$$\tilde{r} = \left[(0.0123)^2 \frac{31}{29} + 2(0.0123)(0.0323) + 0.0323)^2 \frac{11}{9} \right] \Bigg/ \left[(0.0123)^2 \frac{31}{(29)^2} + (0.0323)^2 \frac{11}{(9)^2} \right]$$
$$= 15.14 \text{ and } [t_{15.14}(1 - 0.05/2)] = [t_{15.14}(1 - 0.05/2) = 2.13.$$

Thus, as illustrated in Fig. 4, from Eq. 7, $1 - \beta = 1 - t^{-1}_{15.14,1.42}[2.13] = 0.26$; this study would have limited power to detect a reduction of 0.30 units of behavior from the intervention. By contrast, a naïve power estimate for a comparison made assuming individuals in the same exposure cluster were independent of each other would falsely estimate the power to be ≈ 1.

One might argue that intraclass correlation ρ_A could be higher in the smaller support groups with 10 individuals than the intraclass correlation ρ_B in support groups with 30 individuals. If this is true, then application of the formula of Eq. 7 incorporating ρ_A to obtain U_A and ρ_B to obtain U_B obtains valid estimates.[15]

DISCUSSION

We present power estimates for unbalanced cluster exposure designs with continuous outcomes using the noncentral t distribution. In both examples (as will be the case in general), use of naïve formulas that ignored intracluster correlation greatly overestimated study power. Many previous sample size estimates for clustered designs use more easily calculated Z distribution and central t approximations of the

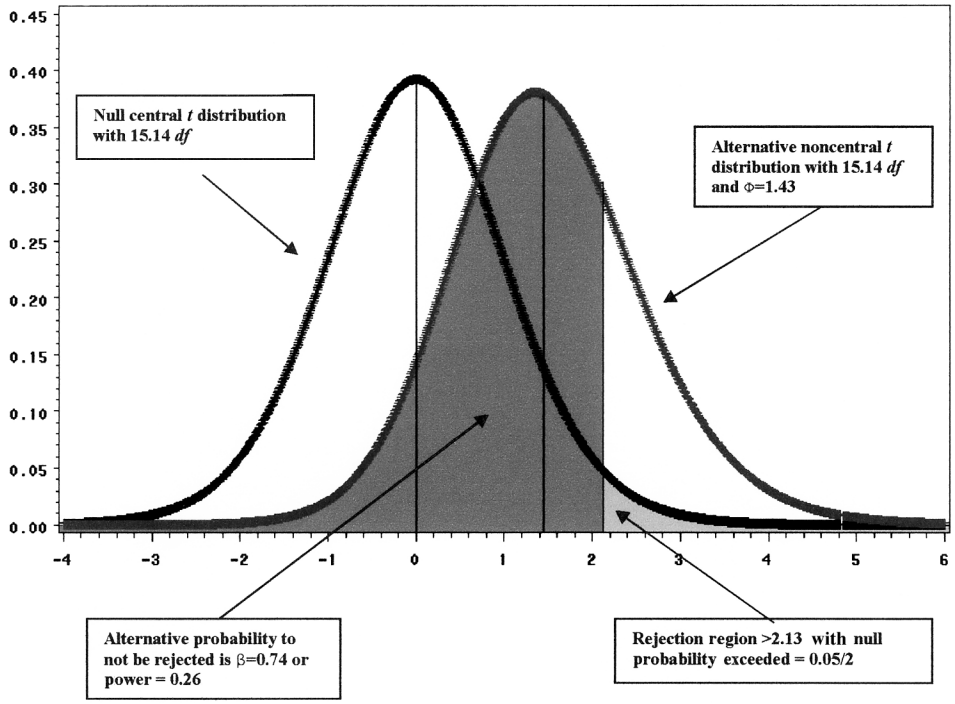

FIGURE 4. Rejection region for t test of example 2 with $r = 15.14$ *df*, $\alpha = .05$, and corresponding type II error under alternative hypothesis with noncentrality parameter $\Phi = 1.43$.

noncentral t distribution. While noncentral t and Z distribution differences are often minor, except perhaps for small values of r, with current computers and software, the need for easy calculation diminishes, and exact answers from noncentral t distributions are feasible.

We considered the intraclass correlation ρ itself to be a nuisance parameter of no interest. In some settings, modeling/estimation of this correlation (or, equivalently, of variance components) may be desired.[10,11,16] While the unbalanced designs complicate such estimation, nested ANOVA models could be used to do this. However, lack of a statistically significant ρ value against $\rho = 0$ should not be used to assume subjects are independent for the purposes of comparing exposure arms. Intraclass correlation estimates often have high variability, and assuming subjects are independent based on such a test of $\rho = 0$ can result in the true type 1 error of the test being much larger than the nominal type 1 error.[15,16]

We gave power estimates for two-sample t-test comparisons between averages of the cluster means in each exposure arm. Another comparison approach that might give different results if $n_A \neq n_B$ would be nested one-way ANOVA models that partition variance. However, this may be difficult in practice as standard ANOVA algorithms require balanced designs ($n_A = n_B$); thus, complicated Sattertwaite[13] adjustments would need to be made. Furthermore, while two-sample t-tests are robust to different intraclass correlations in exposure arms, adjusted nested ANOVAs would not be. While the power of such nested ANOVAs would be difficult to derive, we believe they would be very close to that of two-sample t tests for most settings. Which specific procedure had better power would be a complicated func-

tion of ρ, n_A, and n_B. If one believes a nested ANOVA comparison improves power over the two-sample t test, power estimates in Eq. 5 could be used to obtain conservative power estimates for that test.

Estimation of Minimal Sample Size and Minimal Detectable Differences

The formulas for power given can be transformed into algorithms to estimate required sample size; Eqs. 5 and 7 can be incorporated into software that iteratively estimates minimal values for n, n_A, n_B, k_A, k_B or those needed to attain given power. But, the large number of parameters and wide range of potential constraints (as discussed below) complicate presentation of universal formulas for sample size and detectable differences. Still, in most settings, repeated implementation of Eqs. 5 and 7 through the code in Appendices 1 and 2 can feasibly find minimal sample sizes or detectable differences that will give a specified power.

Let us return to example 1 with $k_A = 100$ high-crime and $k_B = 300$ low-crime neighborhoods, between-person standard deviation $\sigma = 20$, within-neighborhood correlation of $\rho = 0.4$, and minimal difference $\Delta = 5$. We need the minimum cluster size n that gives at least 80% power. Repeated application of Eq. 5 with the code of Appendix 1 finds that $n = 3$ provides 79.7% power, while $n = 4$ provides 83.0% power. We can also return to example 2, with $n_A = 10$ and $n_B = 30$ subjects in an intervention arm for treatments A and B, respectively; between-person standard deviation of $\sigma = 1$; and within-intervention correlation of $\rho = 0.3$. Assuming it is necessary to allocate equal numbers of subjects to each treatment and it is necessary to have ~80% power, repeated application of Eq. 7 finds that 1,200 subjects in each arm (or 2,400 total) (k_A= 120 and $k_B = 40$) gives 79.9% power.

Similar repeated applications of Eqs. 5 and 7 can find minimal values of Δ (to a reasonable number of significant digits) giving a desired power. In examples 1 and 2, respectively (with the original numbers of subjects and clusters presented in the Introduction), the minimal values of Δ to two significant digits that can be detected with 80% power are $\Delta = 4.4$ for example 1 and $\Delta = 0.64$ for example 2.

Implications for Study Design

Optimization by Increasing $\mathbf{k_A}$ *and* $\mathbf{k_B}$ *and Having* $\mathbf{k_A = k_B}$ If $\rho > 0$, then from both Eq. 5 and Eq. 7, given N is fixed, power is maximized by making k_A and k_B as large as possible. This minimizes noncentrality parameters $\Phi = \Delta/\sigma\sqrt{(1/k_A + 1/k_B)(1/n + \rho(1 - 1/n))}$ and $\tilde{\Phi} = \Delta/\sqrt{U_A + U_B} = \Delta/\sqrt{\sigma^2(1/n_A + \rho(1 - 1/n_A))/k_A + \sigma^2(1/n_B + \rho(1 - 1/n_B))/k_B}$. With equal n, in both exposure arms, balanced clusters (with $k_A = k_B$) maximizes power again by maximizing the noncentrality parameter. However, sometimes balanced allocation is not possible or cost-effective.

Optimization of Power Through Balance of Study Subjects If N, k_A, and k_B are fixed, then for $\rho < 1$, the optimal choice for n_A and n_B to maximize power is one that balances the number of subjects across arms with $n_A k_A = n_B k_B$. This follows as the variance of the arm means difference $\rho(1/k_A + 1/k_B) + (1 - \rho)(1/(k_A n_A) + 1/(N - n_A k_A))$ is minimized when $n_A k_A = n_B k_B$. In example 1, with 300 subjects from 100 high-crime neighborhoods compared to 900 from 300 other neighborhoods ($n = 3$ subjects/neighborhood), the power to detect a difference of $\Delta = 5$ when $\sigma = 20$, $\rho = 0.4$, and $\alpha = .05$ from Eq. 5 was 79.9%. Had the 1,200 (300 + 900) subjects been allocated equally to each arm with 600 in low-crime neighborhoods ($n_A = 6$) and

600 in other neighborhoods ($n_B = 2$), then from Eq. 7, the power to detect this difference with the same Δ, σ, and α increases to 82.8%.

Optimal Allocation of Clusters When $\mathbf{n_A \neq n_B}$ If by design $n_A \neq n_B$, an approximate optimal allocation to minimize total number of subjects needed occurs when $k_A \approx k_B(n_B/n_A)\sqrt{(1 + \rho(n_A - 1))/(1 + \rho(n_B - 1))}$ which results in $\sqrt{(1 + \rho(n_A - 1))/(1 + \rho(n_B - 1))}$ subjects in arm A for every subject in arm B. In example 2, with $n_A = 10$ and $n_B = 30$, this corresponds to $k_A = 1.61(k_B)$. While making k_A exactly 1.61 times as large as k_B might be difficult in practice, if 1.5 times as many subjects were allocated to exposure arm B than to exposure arm A, then from Eq. 7 only 2,250 total subjects with $k_A = 90$ (for 900 subjects total in A) and $k_B = 45$ (for 1,350 subjects in B) will give 79.9% power. This compares to 2,400 subjects needed to obtain comparable power with 1,200 subjects in each arm, as was shown previously.

Cost-Beneficial Bound to Number of Subjects per Cluster For fixed numbers of clusters in each treatment arm k or (k_A and k_B), a cost-beneficial upper limit for the number of subjects in each cluster n may be $(5/\rho) - 5$. The standard deviation of arm differences is proportional to $\sqrt{\rho + (1 - \rho)/n}$ and bounded below by $\sqrt{\rho}$. If ρ^2 is much smaller than ρ, then with $n = (5/\rho) - 5$, $\sqrt{\rho + (1 - \rho)/n} \approx (1.1\sqrt{\rho})$ which is close to $\sqrt{\rho}$, the lower bound for variance. Even huge increases in n will not lower the standard deviation much. In example 1 with $\rho = 0.4$, an upper limit for n by this paradigm would be ~8. For $\rho = 0.1$, the upper limit for n would be ~46. In other words, if the intraclass correlation is $\rho = 0.4$, increasing the number of subjects per cluster beyond 8 may not be beneficial from a power standpoint. If $\rho = 0.1$, then increasing the number of individuals per cluster beyond 46 may not be beneficial from a power standpoint. Figure 5 illustrates these points and the effect

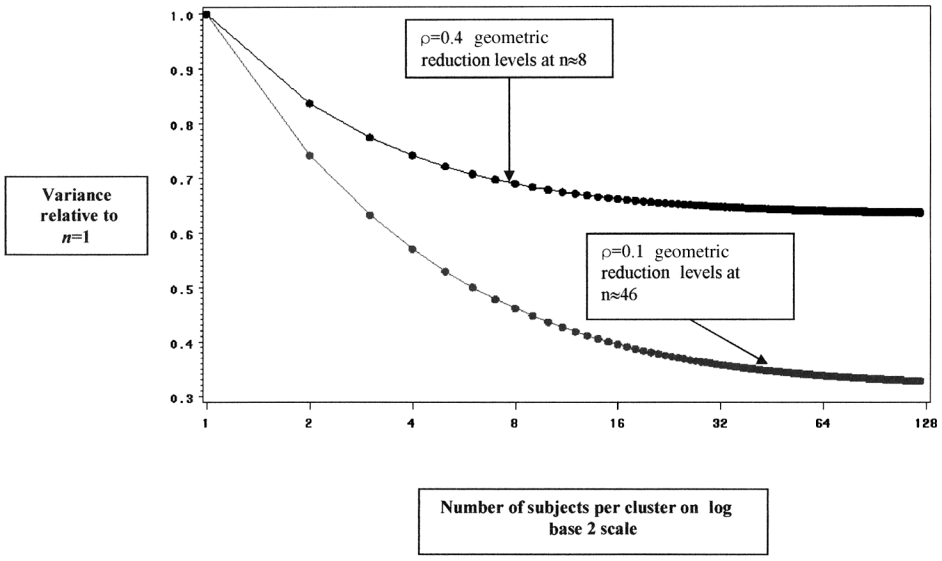

FIGURE 5. Relative variance of comparison of groups with respect to number of subjects per cluster.

on standard deviation of comparisons from increasing n in a study with $\rho = 0.4$ and $\rho = 0.1$. If ρ^2 is not much smaller than ρ (i.e., when ρ is close to 1), then the cost-benefit variance reduction limit occurs with n even smaller than $(5/\rho) - 5$. If costs of sampling clusters and individuals are available, optimal numbers of clusters and individuals to sample could be obtained by extending approaches used for balanced designs.[10,11]

Futility Limit to Number of Subjects per Cluster If the number of clusters k_A and k_B are fixed, then $\Phi' = \Delta/\sigma\sqrt{\rho(1/k_A + 1/k_B)}$ is a lower bound on the noncentrality parameter that will not be exceeded no matter how large n (or n_A and n_B) are made. Thus, an upper bound on power from increasing number of subjects per cluster is $1 - \beta$, where $\beta = t^{-1}_{r,\Phi'}[t_r(1 - \alpha/2)]$. Returning to example 1, $\Phi' = 5/20\sqrt{0.4(1/100 + 1/300)} = 3.42$. Thus, if testing is at $\alpha = .01$, an upper limit for power is $1 - t^{-1}_{398,3.42}[t_{398}.995] = 0.80$, or in other words, no matter how many subjects per cluster are recruited, it is impossible to attain 90% power.

Uneven Numbers of Subjects per Cluster Sometimes, adjustments may be needed due to uneven cluster sample sizes. Often, it is not possible to maintain the same number of individuals in each cluster within study groups, that is, to hold n or n_A and/or n_B constant. This could occur due to limited subjects available for some clusters, nonresponse, or other data loss. One approach, extending an idea suggested by Donner et al.,[4] is to input $\bar{n}$ into Eq. 5 or $\bar{n}_A$ and $\bar{n}_B$ into Eq. 7, where these are the average sizes across the k, k_A, and k_B clusters. Should the cluster sizes vary greatly within the exposure arms, then it may be better to use anticipated harmonic means, $\bar{n}^H_A$ in Eq. 5 or $\bar{n}^H_A$ and $\bar{n}^H_B$ in Eq. 7.[15] Since the exact cluster sizes would likely be unknown in advance and would require numerous terms, such approximations to Eqs. 5 and 7, rather than power formulas incorporating specific cluster sizes, are probably more feasible for the variable cluster size setting.

Complex Samples In many settings (particularly public use data sets), hierarchical and/or weighted sampling schemes were used to select the sample. Appendix 3 describes why sampling weights may not need to be considered and how to estimate ρ to input into Eqs. 5 and 7 in these settings.

Variability of Exposure Within Cluster The methods of this article require all members of the cluster to have the same exposure status. Multilevel models might be needed to analyze cluster studies with exposure that varies within cluster on an individual basis.[10,11] While specific power estimation approaches have not yet been presented for clusters of individuals with different exposure status, they seem feasible by extending methods for stratified comparisons.

SUMMARY

Unbalanced cluster study designs are becoming more common in health outcomes and epidemiologic research, but tools for power/sample size estimation and optimal design are limited. The formulas of Eqs. 5 and 7 estimate power for continuous outcomes in cluster exposure studies with unequal numbers of clusters and/or unequal numbers of subjects per cluster in each arm. Software to implement these algorithms is given in Appendices 1 and 2. From iterative application of Eqs. 5 and 7, minimal sample size and minimal detectable difference can be obtained. But Eqs.

5 and 7 may not directly apply to binomial outcomes. As Appendix 3 describes, Eqs. 5 and 7 may extend to testing of certain types of hypotheses in hierarchical weighted samples. When feasible, power is generally optimized by having the same number of clusters in each arm and (irrespective of numbers of clusters in each arm) the same total number of subjects in each arm. For cost-effectiveness, the upper limit for numbers of subjects per cluster may be $n = 5/\rho - \rho$ or smaller.

APPENDIX 1

SAS Program to Implement the Power Formula of Equation 5

```
/*********************************************************************
Program to Calculate Power for Unbalanced Comparisons for two Groups
with Clustered Observations having Equal Numbers of Subjects per
Cluster (n) and Unequal Numbers of Clusters per Arm (k_a and k_b).

Note - Two Sided Hypothesis Testing is Assumed.  For one Sided Tests,
Double the Value of Alpha Input into Formula.

*********************************************************************/
data a;

input alpha delta stddev n k_a k_b rho;

/Values From Example One are Used                                    */
cards;

0.05 5 20 10 100 300 0.4

data; set a;
label alpha   = 'Type I Error, Two Sided'
      delta   = 'Difference Under Alternative Hypoth'
      stddev  = 'Standard Deviation of One Observation'
      n       = 'Number of Observations per Cluster'
      k_a     = 'Number of Clusters in Arm A'
      k_b     = 'Number of Clusters in Arm B'
      rho     = 'Correlation Between Obs in same Cluster'
      df      = 'Degrees of Freedom'
      tr_null = 'Upper Rejection Limit of Null Hypothesis'
      noncent = 'Noncentrality of Alternative Hypothesis'
      power   = 'Power to Detect the Alternative';

/* Next Line Calculates Degrees of Freedom                          */
df = k_a + k_b - 2;

/*Next Line Calculates Rejection Region of Null
                                          Hypothesis (t_1-a/2,df)*/
tr_null= tinv(1-alpha/2,df);

/* Next Line Calculates Noncentrality Parameter (Theta)             */
noncent = delta/(stddev*((1/k_a+1/k_b)*(1/n+rho*(1-1/n)))**0.5);

/* Next Line Estimates Power According to Formula (5)               */
power=1 - cdf("T",tr_null,df,noncent);

proc print;
var alpha delta stddev n k_a k_b rho power;

run;
```

APPENDIX 2

SAS Program to Implement the Power Formula of Equation 7

```
/*********************************************************************
Program to Calculate Power for Unbalanced Comparisons for two Arms with
Clustered Observations having Unequal Numbers of Subjects per Cluster
(n_a and n_b) and/or Unequal Numbers of Clusters per Arm (k_a and k_b).

Note - Two Sided Hypothesis Testing is Assumed.  For one Sided Tests,
Double the Value of Alpha Input into Formula.
*********************************************************************/
data a;

input alpha delta stddev n_a n_b k_a k_b rho;

/Value From Example Two are Used                                    */
cards;

0.05 0.3 1 30 10 10 30 0.4

data; set a;
label alpha   = 'Type I Error, Two Sided'
      delta   = 'Difference Under Alternative Hypoth'
      stddev  = 'Standard Deviation of One Observation'
      n_a     = 'Number of Obs in Arm A Cluster'
      n_b     = 'Number of Obs in Arm B Cluster'
      k_a     = 'Number of Clusters in Arm A'
      k_b     = 'Number of Clusters in Arm B'
      rho     = 'Correlation Between Obs in same Cluster'
      df      = 'Degrees of Freedom'
      tr_null = 'Upper Rejection Limit of Null Hypothesis'
      noncent = 'Noncentrality of Alternative Hypothesis'
      power   = 'Power to Detect the Alternative';

/* The Next Two Lines Calculate DiSantostefano and Muller Parameters
for Noncentrality and Degrees of Freedom Estimate                   */
u_a = (stddev)**2*(1/n_a + rho*(1-1/n_a))/k_a;
u_b = (stddev)**2*(1/n_b + rho*(1-1/n_b))/k_b;

/*The Next Line Calculates Numerator of Degrees of Freedom Estimate*/
df_num = u_a**2*(k_a+1)/(k_a-1)+2*u_a*u_b+u_b**2*(k_b+1)/(k_b-1);

/*The Next Line Calculates Denominator of Degrees of Freedom Estimate*/
df_den = u_a**2*(k_a+1)/((k_a-1)**2)+u_b**2*(k_b+1)/((k_b-1)**2);

/* Next Line Calculates Degrees of Freedom                           */
df=df_num/df_den;

/*Next Line Calculates Rejection Region of
                                        Null Hypothesis (t_1-a/2,df)*/
tr_null= tinv(1-alpha/2,df);

/* Next Line Calculates Noncentrality Parameter (Theta)             */
noncent = delta/(u_a+u_b)**0.5;

/* Next Line Estimates Power According to Formula (7)               */

power=1 - cdf("T",tr_null,df,noncent);

proc print;
var alpha delta stddev n_a n_b k_a k_b rho power;

run;
```

APPENDIX 3

Application to Complex Sampling Schemes

Many cluster samples used in health outcomes and epidemiologic research are obtained from complex weighted sampling schemes with hierarchical levels of clusters. Examples include large public use data sets such as the Medicare Beneficiary Survey (cf. Olin et al.[17]). In some settings, the formulas of Eqs. 5 and 7 can be applied to complex hierarchical weighted samples.

Hierarchical Clustering Sometimes multiple (hierarchical) levels of clustering exist. We could, for example, consider a study of violence witnessed by schoolchildren in rural and urban areas. The study has two levels of clusters; 150 ZIP codes are randomly chosen, $k_A = 50$ from urban areas and $k_B = 100$ from rural areas. Within each ZIP code, two schools are randomly chosen (schools are clustered within ZIP code), and three students are randomly selected from each school (students are clustered within school). A standardized survey instrument is used to obtain a "violence witnessed" score from each student. ZIP code is the first level of clustering, and school is the second level. By convention, the first level of clustering is denoted the "primary sampling unit." Due to the lower levels of clustering, some pairs of observations in a primary sampling unit have different correlations with each other. In the previous study, among the six students in the same ZIP code, those from the same schools have a higher correlation with each other than do those from different schools.

Let ρ' be the average correlation between two observations in a primary sampling unit. Then (no matter how many levels of sampling), due to the concept of exchangeability, the formulas of Eqs. 5 and 7 can be used with ρ' substituted for ρ. While ρ' may be difficult to estimate, sometimes it can be calculated from available information. In the previous study, let the correlation of violence witnessed between two students in different schools in the same ZIP code be 0.1 and the correlation between two students from the same school be 0.5. Then, since for each student two of the other five students in the primary sampling unit (ZIP code) will be from the same school, $\rho' = 0.26 = 2/5*0.5 + 3/5*0.1$.

Weighted Sampling In many cluster sampling schemes, individuals are not sampled with the same probability. Consider example 1 of this article, with 10 individuals chosen from 400 neighborhoods (100 low crime and 300 high crime). Since neighborhoods have different sizes, individuals from smaller neighborhoods have a greater probability to be sampled than do those from larger neighborhoods. For example, if neighborhood I has 1,000 persons and neighborhood I' has 10,000 persons, then the probability for a person to be sampled in I is $f_I = 0.01$, or 10 times greater than $f_{I'} = 0.001$, the probability for a person to be sampled in I'.

Sometimes, sampling weights w_i are assigned to cluster i to obtain weighted means that adjust for this imbalance. If, in example 1, one wanted to obtain the average cognitive score of all persons living in low- and high-crime neighborhoods, this would be computed as $\overline{X}_{A,W} = \dfrac{\sum_i w_{A,i}\overline{X}_{A,i}}{\sum_i w_{A,i}}$ and $\overline{X}_{B,W} = \dfrac{\sum_i w_{B,i}\overline{X}_{B,i}}{\sum_i w_{B,i}}$, where $w_{A,i}$ and $w_{B,i}$ are the number of people in the respective neighborhoods. So long as the sampling weights w_i are assigned at the cluster level and the tests concern hypotheses of

neighborhood level effects, we argue that sampling weights do not need to be included, making Eq. 2 a valid test statistic and Eqs. 5 and 7 valid power estimators.

Under this design, for each cluster in A, $\overline{X}_{A,i} = u_A + \gamma_{A,i} + \bar{\varepsilon}_{A,i}$, where $\gamma_{A,i}$ is the effect that the ith cluster and has expectation 0 and variance $\rho\sigma^2$ while $\bar{\varepsilon}_{i,A}$ is the mean of individual errors and has expectation zero and variance $(1 - \rho)\sigma^2/n_A$. A similar representation holds for each cluster in B. Since estimation of u_A is the goal, Eq. 2 provides the best estimate with minimal variance. Unless $\gamma_{A,i}$ (and $\gamma_{B,i}$) are correlated with neighborhood size, $w_{A,i}$ (and $w_{B,i}$) are not needed in the analysis. If $\gamma_{A,i}$ (and $\gamma_{B,i}$) are correlated with neighborhood size, then neighborhood size should be adjusted for as a covariate in multivariate analysis rather than through inclusion of sample weights. The methods described in this article, however, do not apply if individual sampling weights vary within cluster and/or direct comparison of weighted arm means is the objective.

ACKNOWLEDGEMENT

Funding for this research was provided by grant P50-MH43450 from the National Institutes of Mental Health. I thank Dr. Sandro Galea for insightful discussion and review of the paper and two anonymous reviewers for helpful comments.

REFERENCES

1. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978;108(3): 100–102.
2. Susser E, Desvarieux M, Wittkowski KM. Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. *Am J Public Health*. 1988; 88(4):671–674.
3. Hansen MH, Hurwitz WN, Madow WG. *Sample Survey Methods and Theory. Volume 1. Methods and Application*. New York: Wiley; 1953.
4. Donner A, Birkett N, Buck C. Randomization by cluster, sample size requirements and analysis. *Am J Epidemiol*. 1981;114(6):906–914.
5. Falconer DS. *Introduction to Quantitative Genetics*. New York: Ronald Press; 1960.
6. Dixon WJ, Massey FM. *Introduction to Statistical Analysis*. New York: McGraw Hill; 1983.
7. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med*. 1988;7(11):1195–1201.
8. Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford, UK: Clarendon Press; 1994.
9. Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. *Health Serv Res Policy*. 2000;5(1):12–16.
10. Snijders TAB, Bosker RJ. Standard errors and sample sizes for two level research. *J Educ Stat*. 1993;18:237–259.
11. Cohen MP. Sample sizes for surveys with data analyzed by hierarchical models. *J Official Stat*. 1998;14(3):267–275.
12. Dunn OJ, Clark VA. *Applied Statistics: Analysis of Variance and Regression*. New York: Wiley and Sons; 1974.
13. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bull*. 1946;2(6):110–114.
14. Di Santostefano RL, Muller KE. A comparison of power approximations for Satterthwaite's test. *Commun Stat Sim*. 1995;24(3):583–593.

15. Hoover DR. Clinical trials of behavioral interventions with heterogeneous teaching subgroup effects. *Stat Med*. In press.
16. Donner A, Klar N. Cluster randomization trials in epidemiology: theory and application. *J Stat Plan Inf*. 1994;42:37–56.
17. Olin GL, Liu H, Merriman B. *Health and Health Care of the Medicare Population. Data From the 1995 Medicare Beneficiary Survey*. Rockville, MD: Westat; 1996.